5.2 An 80-Tile 1.28TFLOPS Network-on-Chip in 65nm CMOS

Sriram Vangal[1,2], Jason Howard[1], Gregory Ruhl[1], Saurabh Dighe[1], Howard Wilson[1], James Tschanz[1], David Finan[1], Priya Iyer[1], Arvind Singh[1], Tiju Jacob[1], Shailendra Jain[1], Sriram Venkataraman[1], Yatin Hoskote[1], Nitin Borkar[1]

[1]Intel, Hillsboro, Oregon, [2]Linköping University, Linköping, Sweden

The ever shrinking size of MOS transistors brings the promise of scalable network-on-chip (NoC) architectures [1] containing hundreds of integrated processing elements with on-chip communication. The NoC architecture (Fig. 5.2.1) contains 80 tiles arranged as a 10×8 2D mesh network and operating at 4GHz. Each tile consists of a processing engine (PE) connected to a 5-port router with mesochronous interfaces, which forwards packets between tiles. The 80-tile on-chip network enables a bisection bandwidth of 256GB/s. The PE contains two independent fully-pipelined single-precision floating-point multiply-accumulator (FPMAC) units, 3KB single-cycle instruction memory (IMEM), and 2KB data memory (DMEM). A 96-bit VLIW encodes up to eight operations per cycle. With a 10-port (6-read, 4-write) register file, the architecture allows scheduling to both FPMACs, simultaneous DMEM load and stores, packet send/receive from mesh network, program control, and dynamic sleep instructions. A router interface block (RIB) handles packet encapsulation between the PE and router. The fully symmetric architecture allows any PE to send (receive) instruction and data packets to (from) any other tile.

The 9-stage pipelined FPMAC architecture (Fig. 5.2.2) uses a single-cycle accumulate algorithm [2] with base 32 and internal carry-save arithmetic and delayed addition. To achieve fast single-cycle accumulate operation, we performed the following three optimizations. First, the accumulator (stage S_5) retains the multiplier output in carry-save format and uses an array of 4-2 carry-save adders to accumulate the result in an intermediate format. This removes the need for a carry-propagate adder in the critical path. Second, accumulation is performed in base 32, converting expensive variable shifters in the accumulate loop to constant shifters. Third, the costly normalization step is moved outside the accumulate loop, where the accumulation result in carry-save is added (stage S_6), normalized (stage S_7) and converted back to base 2 (stage S_8). Careful pipeline re-balancing and a 15FO4 design allow removal of 3 pipe-stages (25%) over work in [2]. This approach reduces latency of dependent FPMAC instructions and enables a sustained multiply-add result (2FLOPS) every cycle. The dual FPMACs in each PE provide 16GFLOPS of aggregate performance.

The 4GHz 5-port wormhole-switched router (Fig. 5.2.3) uses two logical lanes for dead-lock free routing, and a fully non-blocking crossbar switch with a total bandwidth of 80GB/s. Each lane has a 16 FLIT (FLow control unIT) queue, arbiter and flow control logic. The router uses a 5-stage pipeline with a two-stage round-robin arbitration scheme that first binds an input port to an output port in each lane and then selects a pending FLIT from one of the two lanes. A shared-datapath architecture allows crossbar switch reuse across both lanes on a per-FLIT basis. In addition, each 36-bit crossbar data bus is double-pumped at the 4th pipe-stage by interleaving alternate data bits using dual edge-triggered flip-flops, reducing crossbar area by 50%. Combined application of both ideas enables a compact 0.34mm^2 design, resulting in a 34% reduction in router area, 26% fewer devices, 13% improvement in average power, and one cycle latency reduction over the design in [3] when ported and compared in same 65nm process [4]. The point-to-point router links implement a phase-tolerant mesochronous interface with FIFO-based synchronization. A 4-deep circular FIFO captures data using the delayed link strobe at the receiver.

The chip uses scalable global mesochronous clocking, which allows for clock-phase-insensitive communication across tiles and synchronous operation within each tile. The on-chip PLL output (Fig. 5.2.4) is routed using horizontal M8 and vertical M7 spines. Each spine consists of differential clocks for low duty-cycle variation along the worst-case clock route of 26mm. An opamp at each tile converts the differential clocks to a single-ended clock with 50% duty cycle. The worst-case simulated global duty-cycle variation is 3ps and local clock skew within the tile is 4ps. Fig. 5.2.4 also shows clock arrival times for all 80 tiles. The systematic clock skews inherent in the distribution help spread clock power due to simultaneous clock switching over the entire cycle. The estimated global clock distribution power at 4GHz, 1.2V supply is 2.2W.

Fine-grained clock gating, sleep transistor and body bias circuits [5] are used to reduce active and standby leakage power, and are controlled at full-chip, tile-slice, and individual tile levels based on workload. Each tile is partitioned into 21 smaller sleep regions with dynamic control of individual blocks in PE and router units, based on instruction type. The router is enabled on a per-port basis, depending on network traffic patterns. The design uses NMOS sleep transistors to reduce frequency penalty and area overhead. Each FPMAC implements a 6-cycle pipelined wakeup sequence (Fig. 5.2.5) that largely mitigates current spikes over a single-cycle re-activation scheme, while allowing FPMAC execution to start one-cycle into wakeup. Memory arrays use active clamped sleep transistor [6] that ensures data retention and minimizes standby leakage power. The closed-loop opamp configuration ensures that the virtual ground voltage (V_{SSV}) is no greater than a V_{REF} input voltage under PVT variations. V_{REF} is set based on memory cell standby V_{MIN} voltage. The average sleep transistor area overhead is 5.4% with a 4% frequency penalty. About 90% of FPMAC logic and 74% of each PE is sleep-enabled. To allow 4GHz operation, critical registers in the FPMAC and router logic use implicit-pulsed semi-dynamic flip-flops (SDFF) [2]. In addition, forward body bias can be applied to NMOS devices during active mode to increase the operating frequency and reverse body bias can be applied during idle mode for further leakage savings.

Several numerical algorithms in LAPACK have been mapped to the design. Simulated chip frequency versus V_{cc} (Fig. 5.2.6) at 110°C shows tile maximum frequency (f_{MAX}) of 3.13GHz at 1V and 4GHz at 1.2V. With all 80 tiles actively performing block-matrix operations (N=80), the chip achieves a peak performance of 1.0TFLOPS at 1V and 1.28TFLOPS at 1.2V. Estimated typical power consumption is 98W at 1V and 181W at 1.2V. Figure 5.2.6 also plots the design energy efficiency in GFLOPS/W with V_{cc} and frequency scaling, allowing up to 27GFLOPS/W and 310GFLOPS of total performance at 0.6V with an estimated power dissipation of 11W.

Chip micrograph and summary are shown in Fig. 5.2.7. Using a fully-tiled approach, each 3mm^2 tile is drawn complete with C4 bumps, power, global clock and signal routing, and arrayed by abutment. The 275mm^2 custom layout has 100M transistors and 8390 C4 solder bumps, attached to a 14-layer organic package. Test and debug features include a TAP controller and full-scan support for all memory blocks on chip.

Acknowledgements:
The authors thank V. Erraguntla, V. De, D. Somasekhar, D. Jenkins, C. Roberts, B. Nefcy, S. Saha, M. Haycock, S. Borkar, J. Schutz, and J. Rattner for help, encouragement, and support; the LTD and ATD teams for PLL and package designs, and entire mask design team for chip layout.

References:
[1] L. Benini, and G. De Micheli, "Networks on Chips: A New SoC Paradigm," *IEEE Computer*, vol. 35, pp. 70–78, Jan., 2002.
[2] S. Vangal, Y. Hoskote, N. Y. Borkar, et al., "A 6.2-GFlops Floating-Point Multiply-Accumulator with Conditional Normalization," *IEEE J. Solid-State Circuits*, pp. 2314–2323, Oct., 2006.
[3] S. Vangal, N. Y. Borkar, and A. Alvandpour, "A Six-Port 57GB/s Double-Pumped Non-blocking Router Core," *Dig. Symp. VLSI Circuits*, pp. 268–269, June 2005.
[4] P. Bai, C. Auth, S. Balakrishnan, et al., "A 65nm Logic Technology Featuring 35nm Gate Lengths, Enhanced Channel Strain, 8 Cu Interconnect Layers, Low-k ILD and 0.57μm^2 SRAM Cell," *IEDM Tech. Dig.*, pp. 657–660, Dec., 2004.
[5] J. Tschanz, S. G. Narendra, Y. Ye, et al., "Dynamic Sleep Transistor and Body Bias for Active Leakage Power Control of Microprocessors," *IEEE J. Solid-State Circuits*, pp. 1838–1845, Nov., 2003.
[6] M. Khellah, N. S. Kim, J. Howard, et al., "A 4.2GHz 0.3mm^2 256kb Dual-Vcc SRAM Building Block in 65nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 624–625, Feb., 2006.

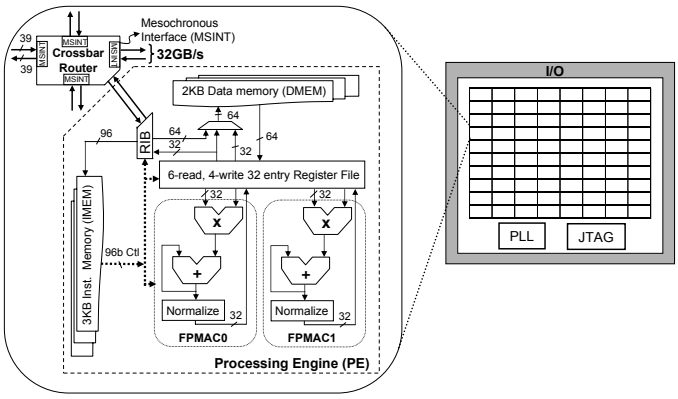


Figure 5.2.1: NoC block diagram and tile architecture.

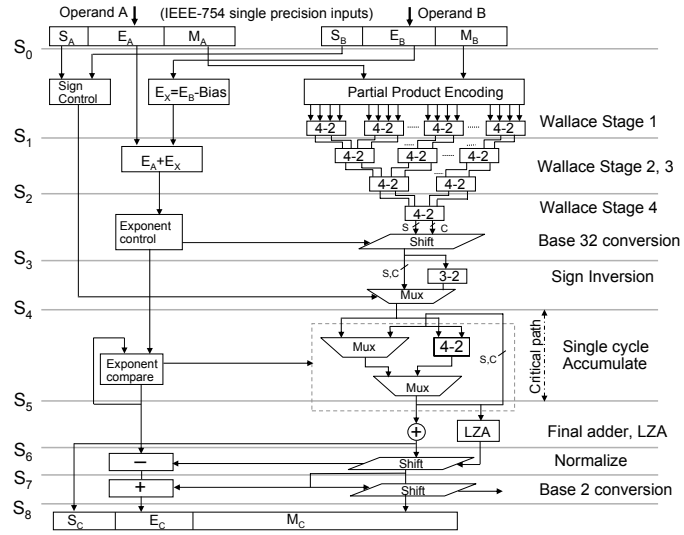


Figure 5.2.2: FPMAC 9-stage pipeline with single-cycle accumulate loop.

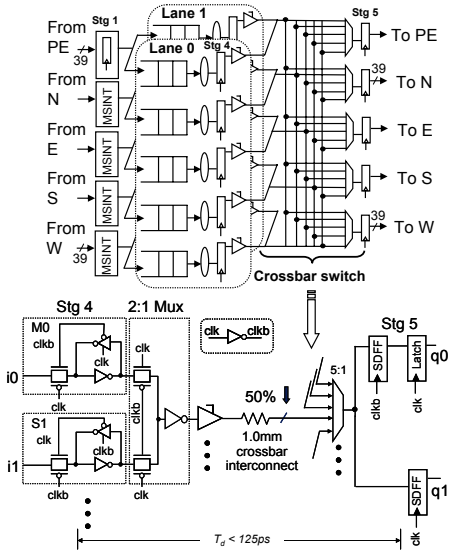


Figure 5.2.3: Shared crossbar router with double-pumped crossbar switch.

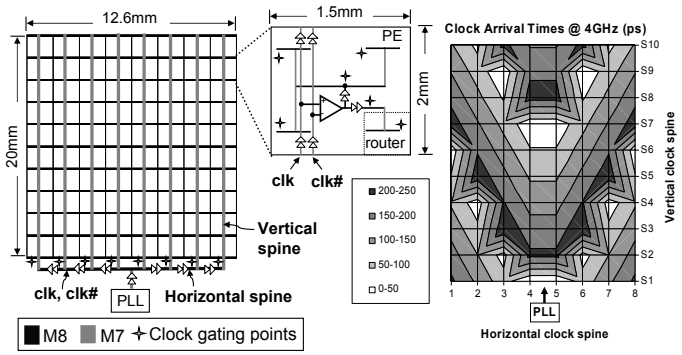


Figure 5.2.4: Global mesochronous clocking and simulated clock arrival times.

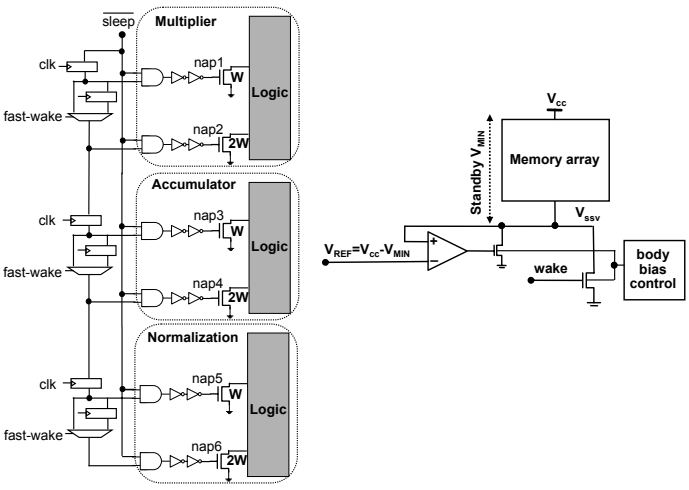


Figure 5.2.5: FPMAC pipelined wakeup diagram and state-retentive memory clamp circuit.

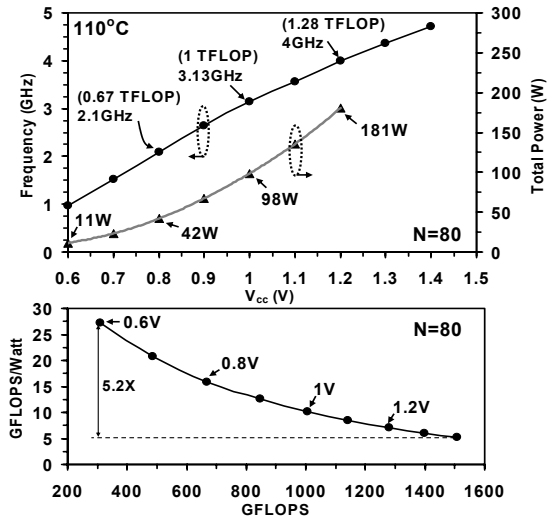
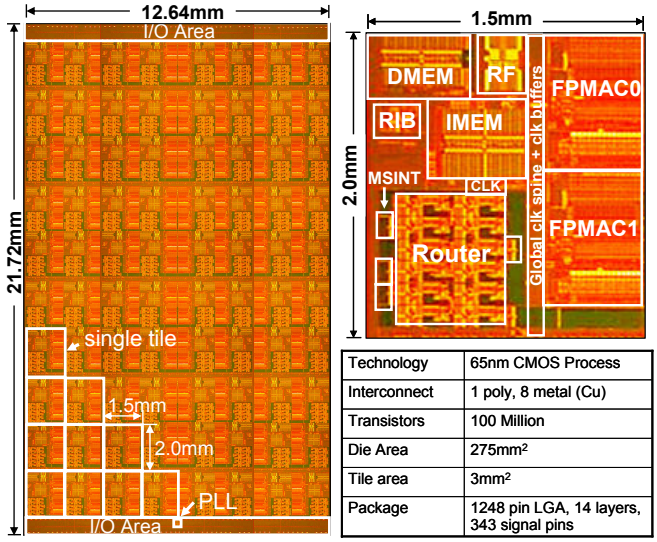


Figure 5.2.6: Estimated frequency and power versus V_{cc}, and power efficiency with 80 tiles (N) active.

Continued on Page 589



Technology	65nm CMOS Process
Interconnect	1 poly, 8 metal (Cu)
Transistors	100 Million
Die Area	275mm^2
Tile area	3mm^2
Package	1248 pin LGA, 14 layers, 343 signal pins

Figure 5.2.7: Full-Chip and tile micrograph and characteristics.

